Föreläsning 10

den 4 maj 2021 12:46

Matematisk Statistik och Diskret Matematik, MVE051/MSG810, VT22

Timo Vilkas

Chalmers - Göteborgs Universitet

Föreläsning 10
3 maj 2022

Skattning av proportioner

Exempel

En studie vill skatta andelen av personer som bor i en viss stad som äger en surfplatta. Ett slumpmässigt valt stickprov på 250 invånare visar att 98 utav dessa angav att äga en surfplatta. Med det i åtanke är $\hat{p} = \frac{98}{250} = 0.392$ en rimlig skattning av proportionen av individer i populationen som äger en surfplatta.

I allmänhet fokuserar vi på en viss egenskap *(eng. trait)* och undersöker om individer i populationen har denna eller inte. Man vill få fram andelen av individer i populationen som har egenskapen, eller närmare sagt en skattning av denna proportion.

- Vi drar ett stickprov av storlek n ur population.
- Låt variablerna $X_1, \ldots, X_n$ indikera om motsvarande individ i stickprovet har egenskapen eller inte, alltså:

$$X_j = \begin{cases} 1 & \text{om individ nummer } j \text{ i stickprovet har egenskapen} \\ 0 & \text{annars} \end{cases}$$

  Enligt våra antaganden är då $X_1, \ldots, X_n$ oberoende Ber(p)-fördelade variabler, där p är den verkliga proportionen i populationen.
- En punktskattning för p är då

$$\hat{p} = \frac{\sum_{j=1}^{n} X_j}{n}.$$

- $\hat{p} = \frac{\sum_{j=1}^{n} X_j}{n}$ är en väntevärdesriktig estimator för p:

$\mathbb{E}\hat{p} \overset{\text{linj.}}{=} \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}X_j = \frac{1}{n}\cdot(np) = p$

$\mathbb{E}X_j = 1\cdot p + 0\cdot(1-p) = p$ för alla $j \in \{1,\dots,n\}$ då $X_j \sim \text{Ber}(p)$.

$\rightarrow$ $\hat{p}$ är en väntevärdesriktig estimator för p.

- Variansen av estimatorn $\hat{p}$:

obs. $X_j^2 = X_j$ för en indikatorvariabel

$Var(X_j) = \mathbb{E}[X_j^2] - \mathbb{E}[X_j]^2 = p - p^2 = p(1-p)$, alltså

$Var(\hat{p}) = \left(\frac{1}{n}\right)^2 \text{var}\left(\sum_{j=1}^{n} X_j\right) \overset{X_j \text{ oberoende och Be}(p)}{=} \frac{1}{n^2}\cdot\sum_{j=1}^{n}\text{var}(X_j) = \frac{n(1-p)p}{n^2} \rightarrow 0$ med $n\rightarrow\infty$

För tillräckligt stort n blir variansen alltså liten, eller med andra ord är $\hat{p}$ en konsistent estimator.

Dessa två egenskaper gör $\hat{p}$ till en bra estimator för p.

KI för proportionen

- Enligt centrala gränsvärdessatsen är $\hat{p}$, för tillräckligt stort n, approximativt normalfördelad med väntevärde p och varians $p(1-p)/n$.
- Ett (symmetriskt) $100(1-\alpha)\%$ konfidensintervall ges därför av

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right),$$

där $z_{\alpha/2}$ som vanligt betecknar $(1-\frac{\alpha}{2})\,100$-percentilen till standard normalfördelningen, dvs. talet så att $\mathbb{P}(Z \leq z_{\alpha/2}) = 1 - \frac{\alpha}{2}$ för $Z \sim \mathcal{N}(0,1)$.

$n = 250$, $\hat{p} = \frac{98}{250} = 0.392$, $\alpha = 0.05$; $z_{\alpha/2} = 1.96$

Exempel

Ett 95% KI för proportionen av invånare som äger en surfplatta (i exemplet ovan) ges därför av

$$\left(0.392 - 1.96\sqrt{\frac{0.392 \cdot 0.608}{250}}, 0.392 + 1.96\sqrt{\frac{0.392 \cdot 0.608}{250}}\right)$$

$$= (0.331, 0.453)$$

Baserat på stickprovet är vi 95% säkra att andelen p i populationen som har en surfplatta ligger inom intervallet (0.331, 0.453).

Stickprovets storlek för att skatta p

- Ibland har man möjlighet att välja stickprovets storlek och vill gärna ha kontroll över längden på resulterande KI.
- Punkterna i konfidensintervallet

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

ligger maximalt

$$d = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ifrån punktskattningen $\hat{p}$.

Stickprovets storlek för att skatta p

- För givet d behöver man alltså ett stickprov av storlek

$$n = \frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{d^2},$$

vilket dock kräver en apriori skattning av parametern p (som vi då kan använda som $\hat{p}$ i formeln).

- Om man inte har någon apriori skattning av p tillgänglig, kan man – för att vara på den säkra sidan – välja

$$n = \frac{z_{\alpha/2}^2}{4d^2},$$

då $x(1-x) = -(x-\frac{1}{2})^2 + \frac{1}{4} \leq \frac{1}{4}$ för alla x.

Exempel

Ett mobilföretag vill ta reda på vilken andel av kunderna som är äldre än 50 år regelbundet skickar text meddelanden. Hur många kunder som är äldre än 50 år skulle företaget inkludera i en kundundersökning för att kunna vara 90% säker att den skattade proportionen (ifrån stickprovet) inte avviker fler än 3% ifrån den verkliga andelen i populationen (av kunder som är 50+) som regelbundet skickar text meddelanden givet att...

1. ... $\hat{p} = 0.4$ är en apriori skattning för p.
2. ... ingen apriori skattning finns tillgänglig.

Lösning: $\alpha = 0.1$, $z_{0.05} = 1.645$ och $d = 0.03$.

1. $n = \frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{d^2} = \frac{(1.645)^2 \cdot 0.4 \cdot 0.6}{0.03^2} = 721,6 \approx 722$
2. $n = \frac{1}{4}\left(\frac{z_{\alpha/2}}{d}\right)^2 = 751,67 \approx 752.$

Hypotesprövning för proportioner

- På samma sätt som vi genomförde hypotesprövning för medelvärdet kan man göra det för en proportion i populationen.
- En rimlig teststatistik är då

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

där p_0 är nollvärdet för p som används i hypoteserna.
- Om n är tillräckligt stort, är teststatistiken approximativt standard normalfördelad.
- Då det handlar om en approximation av binomialfördelningen med normalfördelningen anses n vara tillräckligt stort om

$$\min\{np_0, n(1-p_0)\} > 5$$

(jfr. med föreläsning 3).

Exempel

Gällande frågan om andelen av nyfödda är större för pojkar än för flickor har tagits fram ett stickprov på 25 468 nyfödda barn.
13 173 av dem var pojkar, vilket ger en stickprovsandel på 0.5172.
Stödjer stickprovet antagandet att det föddes fler pojkar än flickor i hela populationen, till signifikansnivån $\alpha = 0.05$?

Lösning:

$$H_0 : p = 0.5 \quad \text{och} \quad H_1 : p > 0.5$$

Eftersom n är tillräckligt stort är teststatistiken, $Z = \frac{(\hat{p}-0.5)\sqrt{25468}}{\sqrt{0.5 \cdot 0.5}}$ approximativt standard normalfördelad.
Det kritiska värdet ges av $z_{0.05} = 1.645$.
Teststatistikens utfall $z = \frac{(0.5172-0.5)\sqrt{25468}}{0.5} = 5.49 > 1.645$ ligger i den kritiska regionen, vilket leder till att H_0 förkastas och man kan konstatera att stickprovet faktiskt bekräftar alternativet, att andelen pojkar är större bland nyfödda.

Jämförelse av två proportioner

Antag att vi har två olika populationer och vi intresserar oss för en viss egenskap hos individerna. Andelen individer som har denna är okänd för båda populationer. Vi vill jämföra populationerna genom att skatta andelen i båda och tolka differensen.

Exempel

En undersökning genomförs som fokuserar på hur många av forskarna regelbundet använder programmet R, både inom ren matematik och sannolikhetsteori/statistik.
Populationer: Forskare inom ren matte och forskare inom sannolikhetsteori/statistik.
Egenskap av intresse: Användning av programmet R.

Punktskattning och KI för differensen av två proportioner

Antag att den verkliga proportionen i population 1 är p_1 och p_2 i population 2.

- Vi tar fram ett stickprov (av storlek n_1 resp. n_2) ur båda populationer (oberoende såklart)
- För varje beräknar vi en punktskattning: $\hat{p}_1$ och $\hat{p}_2$.
- En skattning för $p_1 - p_2$ blir då $\hat{p}_1 - \hat{p}_2$.
- För tillräckligt stora stickprov är $\hat{p}_1 - \hat{p}_2$ approximativt normalfördelad med väntevärde $p_1 - p_2$ och varians $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$.
- Ett symmetriskt $100(1-\alpha)\%$ KI för $p_1 - p_2$ ges därför av

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Exempel

Vi tar stickprov i exemplet ovan av storlek $n_1 = n_2 = 375$ ifrån båda populationer. Antalet forskare i stickprovet som regelbundet använder R är 195 för population 1 (ren matte) och 232 för population 2 (sanno/statistik). Vi får alltså $\hat{p}_1 = \frac{195}{375} = 0.52$ och $\hat{p}_2 = \frac{232}{375} = 0.619$.
En punktskattning för differensen $p_1 - p_2$ ges då av $0.52 - 0.619 = -0.099$ och dess standardavvikelse är

$$\sqrt{\frac{0.52 \cdot 0.48}{375} + \frac{0.619 \cdot 0.381}{375}} \approx 0.036.$$

Exempel

Ett 95% konfidensintervall för $p_1 - p_2$ ges av

$$(0.52 - 0.619 - 1.96 \cdot 0.036, 0.52 - 0.619 + 1.96 \cdot 0.036),$$

alltså

$$(-0.17, -0.028).$$

Då intervallet inte innehåller 0 och ligger vänster om 0, kan man konstatera med en säkerhet på 95% att andelen av forskarna som regelbundet använder R är större inom population 2 jämfört med population 1.